# Reinforcement Learning and Its Applications in Modern Power and Energy Systems

*A Review*

Cao, Di; Hu, Weihao; Zhao, Junbo; Zhang, Guozhou; Zhang, Bin; Liu, Zhou; Chen, Zhe; Blaabjerg, Frede

*Published in:*
Journal of Modern Power Systems and Clean Energy

*DOI (link to publication from Publisher):*
10.35833/MPCE.2020.000552

*Creative Commons License*
Unspecified

*Publication date:*
2020

*Document Version*
Publisher's PDF, also known as Version of record

# Reinforcement Learning and Its Applications in Modern Power and Energy Systems: A Review

Di Cao, Weihao Hu, Junbo Zhao, Guozhou Zhang, Bin Zhang, Zhou Liu, Zhe Chen, and Frede Blaabjerg

***Abstract*—With the growing integration of distributed energy resources (DERs), flexible loads, and other emerging technologies, there are increasing complexities and uncertainties for modern power and energy systems. This brings great challenges to the operation and control. Besides, with the deployment of advanced sensor and smart meters, a large number of data are generated, which brings opportunities for novel data-driven methods to deal with complicated operation and control issues. Among them, reinforcement learning (RL) is one of the most widely promoted methods for control and optimization problems. This paper provides a comprehensive literature review of RL in terms of basic ideas, various types of algorithms, and their applications in power and energy systems. The challenges and further works are also discussed.**

***Index Terms*—Reinforcement learning, deep reinforcement learning, power system operation and control, optimization.**

## I. Introduction

With the gradual depletion of fossil energy and increasing environmental pressure, a revolution in energy sector is going on globally [1]. This revolution has the following characteristics: high penetration of renewable energies, wide application of power electronic devices, and increasing connection of flexible load, i.e., electric vehicle and distributed energy storage system. These characteristics increase the system complexities and uncertainties, and bring great challenges to the operation of power and energy systems [2]. The physical model based approaches require the accurate mathematical models and parameters, the construction of which is challenging considering the increasing system complexities and uncertainties. With the proliferation of the advanced sensor and smart meters, smart grid is producing data with huge volumes, mutual correlations, and complex structures. The data contain valuable information that can not only be utilized to extract intelligence for operation and planning of power and energy systems, but also complement the shortcomings of the physical model based methods [3]. In the area of big data, machine learning (ML) can help overcome the aforementioned limitations by directly learning from data [4]. They can extract powerful knowledge from historical data to deal with the highly uncertain system dynamics. The learned model is adaptive and can be generalized to newly encountered situations. Among the ML family, reinforcement learning (RL) may be the most suitable one for the optimization and control problems [5].

Various approaches have been proposed for the optimization and control of modern power and energy system. In general, the optimization method can be broadly classified into classical algorithms [6], [7] and heuristic algorithms [8]. Classical algorithms such as stochastic programming and robust optimization are proposed to address uncertainty problems. Those methods deal with uncertainties by finding a pre-determined solution. However, with the increasing penetration of distributed energy resources (DERs) and flexible demands, both the generation and demand sides are facing growing uncertainties. In this context, real-time control strategy based on the latest observation may achieve a better performance than the pre-determined ones. In addition, these methods are based on physical models, which requires accurate physical models for the optimization of a defined objective function. But it is difficult to maintain the reliable physical models in practice, which limits their applicability. Heuristic methods, such as particle swarm optimization, are easy to be implemented, but the computational burden rises exponentially with the number of control variables. For the control of power and energy systems, the pole-placement [9] and residue methods [10] are widely used. However, they only consider the single operation condition of the system, limiting its robustness in other conditions. Therefore, robust method is introduced, such as the fuzzy control [11], $H_\infty$ and $H_2$ control [12], and gap decision theory [13]. However, the solutions obtained by these robust methods are usually conservative. In this context, adaptive control theory is introduced. For example, the synergetic control theory [14], model predictive control [15], and adaptive dynamic programming [16] are proposed. Also, they are the physical model based methods and can be significantly affected by model in-

Manuscript received: July 31, 2020; accepted: October 28, 2020. Date of CrossCheck: October 28, 2020. Date of online publication: November 26, 2020.

This work was supported by the Sichuan Science and Technology Program (Sichuan Distinguished Young Scholars) (No. 2020JDJQ0037).



D. Cao, W. Hu (corresponding author), G. Zhang, and B. Zhang are with the Wide-area Measurement and Control Sichuan Provincial Key Laboratory, School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China (e-mail: caodi@std. uestc. edu. cn; whu@uestc.edu.cn; zgz@std.uestc.edu.cn; sven@uestc.edu.cn).

J. Zhao is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, Mississippi, USA (e-mail: junbo@ece. msstate.edu).

Z. Liu, Z. Chen, and F. Blaabjerg are with the Department of Energy Technology, Aalborg University, Aalborg, Denmark (e-mail: zli@et.aau.dk; zch@et.aau.dk; fbl@et.aau.dk).

DOI: 10.35833/MPCE.2020.000552

accuracies.

Different from the methods mentioned above, RL is a class of method that is inspired from behavioral psychology. RL can extract optimal operational knowledge from historical data through continuous interactions with the environment while the global optimum is unknown. They can get rid of the dependency on the accurate physical model by learning a surrogate model [17] or batch RL [18]. The learned strategy is scalable, thus it can be exploited in an online manner to inform decisions based on the latest information. Owing to these advantages, RL has been widely applied in industrial manufacturing [19], operation and scheduling [20], robotic control [21], etc. There are also wide-area applications in power and energy system, including optimization of smart power and energy distribution grid, demand side management, electricity market, operational control, etc. This paper summarizes the recent researches on RL for the optimization and control of power and energy systems and discusses the potential research directions. The main contributions are as follows.

1) Typical RL, deep RL (DRL), and multi-agent DRL (MADRL) for optimization and control of modern power and energy systems are summarized thoroughly with the detailed analysis of advantages and disadvantages.

2) State-of-the-art applications of RL algorithms in power and energy systems are organized with several categories.

3) A comprehensive analysis of the limitations of current RL algorithms is presented.

The structure of this paper is as follows. Section II introduces the RL algorithms. A comprehensive review of the RL for power systems applications is shown in Section III. Section IV discusses the challenges and prospects of RL in power systems and conclude this paper.

## II. Review of RL Algorithm

In this section, the Markov decision process (MDP) is first illustrated, followed by the classical RL, advanced DRL, and MADRL algorithms.

### A. RL

ML algorithms can be classified into three categories: unsupervised learning, supervised learning, and RL. Unsupervised learning typically includes clustering, dimensionality reduction, and association rule learning methods, etc. Supervised learning, which typically acts as the function approximator, aims to build an affine rule mapping from the training input to the labeled output utilizing a predefined evaluation index [22]. Among supervised learning, deep neural network (DNN) is the one that has attracted more attention in recent years.

Compared with supervised learning and unsupervised learning, RL is regarded as active learning. The basic structure of RL is shown in Fig. 1.

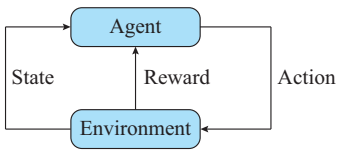


Fig. 1. Framework of RL algorithm.

There are mainly two components: the agent and the environment [23]. The decision process of RL can be described as follows. At each time step, the agent obtains an observation of the environment. Then, it makes decisions according to the observation following the current policy. The environment is affected by the action and then transfers to a new state. Meanwhile, it returns a reward value to the agent for the judgement of the action. The agent aims to maximize the reward obtained from the environment by learning an optimal control strategy through continuous interactions with the environment. RL can develop the optimal control behavior through continuous interaction with the environment and the gradient calculated by the feedback reward signal [23]. The main RL algorithms and their relationship are shown in Fig. 2.

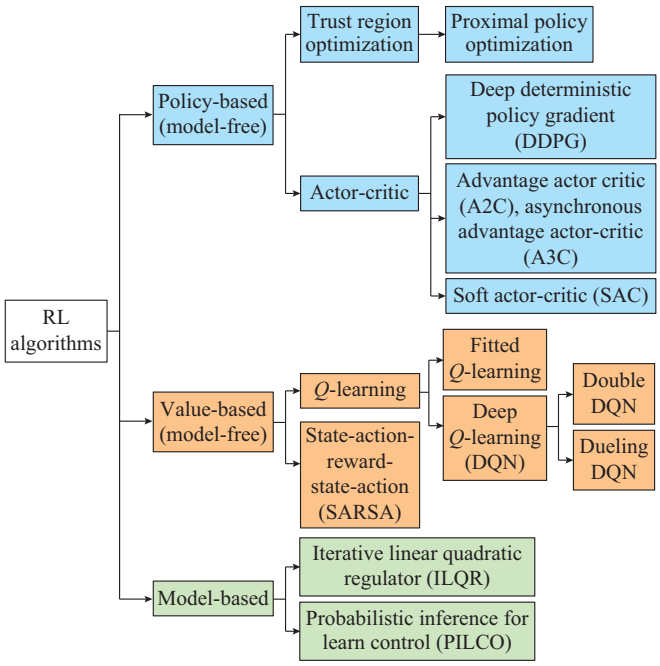


Fig. 2. Main RL algorithms and their relationships.

#### 1) MDP

In the framework of RL, the interaction between agent and environment is formalized by MDP [23]. An MDP can be described by the tuple $\langle s, a, r, T, \gamma, \pi \rangle$, which is explained as follows.

1) Action $a \in A$: $A$ is the action set, and $a$ is a specific action.

2) State $s \in S$: $S$ is a finite state set, and $s$ is a given state.

3) Transition model $T(s, a, s')$~$\Pr(s'|s, a)$: the transition model determines the prediction probability of the next step state $s'$ given the current state $s$ and action $a$.

4) Reward function $r(s, a)$: $r$ is the immediate reward by the agent when taking action $a$ under state $s$.

5) $\gamma \in [0, 1]$: the discount factor $\gamma$ is used to balance the importance of immediate rewards relative to future rewards.

6) Policy $\pi(s) \rightarrow a$: a policy mapping from states to actions is yielded when solving an MDP. An optimal policy $\pi^*$ means that the maximum expected discount cumulative reward can be obtained.

The illustration of MDP is shown in Fig. 3.

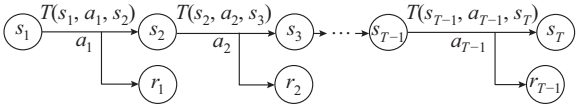


Fig. 3. Illustration of MDP.

At each epoch, the environment takes the current state $s_t$ and action $a_t$ as the input, and the output is the current reward $r_t$ and the state of the next step $s_{t+1}$. The quality of action $a_t$ under state $s_t$ is measured by the cumulative discounted reward, which is obtained by the agent from current time-step onward:

$$Q^{\pi}(s_t, a_t) = E(R_t | s_t = s, a_t = a) = E\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s, a_t = a\right) \quad (1)$$

where $E(\cdot)$ is the expectation of the cumulative discounted reward; $R_t$ is the current cumulative reward by the agent from time step $t$ onward; and $Q^{\pi}(s_t, a_t)$ is the so-called action-value function. RL algorithm aims to look for an optimal policy $\pi^*$ so as to maximize the action-value function.

*2) Q-learning*

Considering that the future system information is unknown, it is intractable for agent to determine the optimal policy $\pi^*$. Thus, iterative update of action-value function based on Bellman equation is adopted by the $Q$-learning algorithm [23] as:

$$Q_{i+1}(s_t, a_t) = E(r_t + \gamma \max Q_i(s_{t+1}, a_{t+1}) | s_t = s, a_t = a) \quad (3)$$

With iteration $i \to \infty$, the $Q$-value will converge to the optimal value $Q^*(s, a)$. Then, the optimal control schedules can be obtained based on a greedy strategy:

$$a^* = \arg\max_{a \in A} Q^*(s, a) \quad (4)$$

Original $Q$-learning algorithm stores the action values in a discretized lookup-table, the size of which is determined by the dimensions of states and actions. However, multivariate continuous state and action variables are typically needed in practical applications of power and energy system. The discretization of the state and action variables not only leads to the sharp increase of the computational complexity, but also wastes valuable information about the structure of state and action domain that are essential for solving problems.

## B. DRL

Traditional RL algorithms have several limitations. Firstly, they suffer from "curse of dimensionality" when coping with scenarios with high-dimension and continuous state and action space. Secondly, hand-specified state representations are typically required. As a function approximator, DNN can be applied to address the above limitations by approximating the state-action function with the parameters of neural network (NN). Combining the DNN and the RL algorithm has two advantages: ① the strong feature extraction ability of DNN helps avoid the manually feature design process, and the control decisions can be directly derived from the raw inputs through end-to-end learning procedure; ② DNN helps RL generalize problems with a large state space [24]. Despite these benefits, there are also some challenges, i.e., the training data of DNN are typically assumed to be independent and identically distributed [25]. However, since RL algorithms are usually applied to solve sequential-control problems, the training data are highly correlated, violating the independence assumption. In addition, the distribution of the training data may be non-stationary when the agent explores the environment, which means the training data may not be identically distributed. The highly correlated and non-stationary training data may cause the divergence of the training process. Various methods have been proposed to address them and they can be classified into two categories: the value-based algorithms and the policy gradient algorithms.

*1) Value-based Algorithm*

One of the breakthroughs for DRL is the value-based DQN algorithm, which uses DNN as the function approximator to fit the action-value function. The structure of the DQN algorithm is shown in Fig. 4.

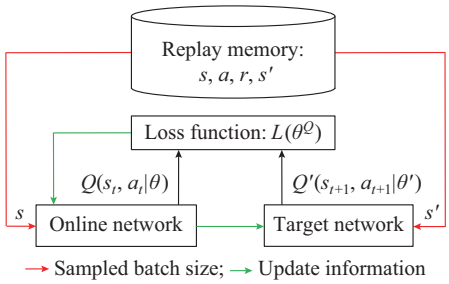


Fig. 4. Structure of DQN algorithm.

DQN algorithm adopts a replay buffer to store a large number of transitions $\langle s, a, r, s' \rangle$. The experience replay mechanism helps break the correlation among training data by randomly sampling a mini-batch data from the memory when updating the NN. DQN also introduces a target $Q$ network to alleviate the non-stationary distribution of training data, significantly improving the stability of training process. At each time step, the parameters of the action-value function are optimized by minimizing the following loss function [26]:

$$L(\theta) = E_{\pi}((Q(s_t, a_t|\theta) - r(s_t, a_t) - \gamma \max_a Q(s_{t+1}, a_{t+1}|\theta'))^2) \quad (5)$$

where $\theta$ and $\theta'$ are the parameters of the action-value function. DQN has several improved versions to reduce overestimation, such as double DQN [27], dueling DQN [28].

The DNN utilized in DQN avoids the discretization of state space. However, since DQN relies on finding an action which maximizes the action-value function, it still needs to discretize the action domain for the applications with continuous action variables. The discretization of action domain may lead to the curse of the dimensionality issue since the number of total actions increases exponentially with the number of action types. Moreover, the discretization of action space may cause information loss and lead to sub-optimal solutions. This makes it intractable to apply the DQN-based method to applications with high-dimension and continuous action space.

*2) Policy Gradient Algorithm*

Policy gradient algorithm is a kind of algorithm suitable for the tasks with continuous and high-dimension action space. Instead of learning the action-value function, policy gradient algorithm directly learns an affine rule mapping

from the observed state to control decision. Policy gradient algorithm maintains a policy function parameterized by the weights of $\theta$. It aims to maximize the expected cumulative reward $E_{x\sim p_\theta(\tau)}(R|\pi)$ by optimizing $\theta$. Specifically, the parameters are optimized via the gradient:

$$\nabla R_\theta = E_{p(\tau|\theta)}(R(\tau)\nabla_\theta \lg p_\theta(\tau)) \tag{6}$$

where $\tau=\{s_1, a_1, s_2, a_2, \ldots, s_T, a_T\}$ is a trajectory, $T$ is the episode length; and $R(\tau)$ is the cumulative reward of the trajectory. The parameters of the NN are optimized towards the direction that increases the probability of the trajectory $\tau$ with a larger reward. The variance of the gradient is high in policy gradient algorithm. To this end, a baseline term is typically subtracted from $R(\tau)$.

The baseline term in the policy gradient algorithms is typically replaced by the value function $V(s)$ learned by the critic function. This fits to the actor-critic algorithm, which is a subset of the policy gradient algorithms. The basic structure of actor-critic-based algorithms is shown in Fig. 5.

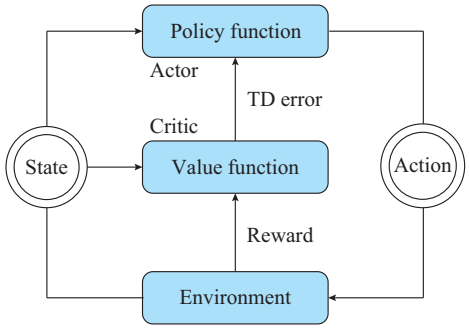


Fig. 5. Structure of actor-critic-based algorithms.

DDPG is an actor-critic-based algorithm. It employs two functions for different purposes: the actor function $\mu(\cdot|\theta^\mu)$ learns the control policy and the critic function $Q(\cdot|\theta^Q)$ provides the judgement of the actor. The actor and critic are trained against each other so that the actor can learn a better control strategy and the critic can provide a more accurate judgement. The DDPG also introduces the experience replay mechanism and the target networks to stabilize the training. The parameters of the critic network $\theta^Q$ are optimized by minimizing the following loss function [29]:

$$\begin{cases} L(\theta^Q) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(Q(s_t, a_t|\theta^Q) - y_t)^2 \\ y_t = r(s_t, a_t) + \gamma Q'(s_{t+1}, \mu'(s_{t+1}|\theta^{\mu'})|\theta^{Q'}) \end{cases} \tag{7}$$

where $N$ is the number of samples in one batch; $Q'(\cdot)$ and $\mu'(\cdot)$ are the target critic and actor networks, respectively; and $\theta^{Q'}$ and $\theta^{\mu'}$ are the parameters of target critic and actor networks, respectively. The parameters of the actor network are optimized according to the following deterministic policy gradient [30]:

$$\nabla_{\theta^\mu}\mu = E\left(\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{s=s_t, u=u(s_t|\theta^\mu)}\right) = E\left(\nabla_\theta \mu_\theta(s|\theta^\mu)|_{s=s_t} \nabla_a Q(s, a|\theta^Q)|_{s=s_t, u=u(s_t|\theta^\mu)}\right) \tag{8}$$

The parameters of target networks are optimized by the soft update mechanism to alleviate the non-stationary distribution of training data.

Different from the experience replay mechanism used in DQN and DDPG, A3C algorithm employs multiple parallelized workers to break the correlations among the training data and stabilize the training. The gradients are first calculated by multiple local actors, and then passed to the global NN to perform the optimization. An entropy term is also added to the loss function to improve exploration and help convergence to a better policy. The parameters of the policy function $\pi(\cdot|\theta^\mu)$ are optimized by [31]:

$$\nabla R_{\theta^\mu} \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}(A(s_t, a_t)\nabla_{\theta^\mu}\lg \pi_{\theta^\mu}(a_t^n|s_t^n) + \beta\nabla_{\theta^\mu} H(\pi(s_t|\theta^\mu)) \tag{9}$$

where $A(s_t, a_t)$ is the advantage function; $H(\cdot)$ is the entropy function; $\beta$ is the weight of the entropy term; and $T$ is the horizon of time step. SAC is also an entropy-regularized-based DRL algorithm. It adopts a value function and two action-value functions, and alternates between updating using the sampled batches from the memory and collecting experiences following the current policy [32].

Classical policy gradient algorithms also include trust region policy optimization (TRPO) [33] and proximal policy optimization (PPO) [34] methods, both of which are proposed to solve the poor data efficiency issues of vanilla policy-based methods. PPO avoids the abrupt policy changes during training by employing a novel objective function with a clipped probability ratio. It achieves better reliability and stability than the vanilla policy gradient algorithms, and is much easier to implement than the TRPO method.

### C. MADRL Algorithm

The algorithms mentioned above only use single agent. However, a lot of applications involve the interactions among multiple agents, such as multiplayer games and multi-robot control problem. The application of single-agent DRL algorithm to multi-agent environment yields a poor performance as the environment can become non-stationary from the point of view of each individual agent. This prevents the use of memory replay mechanism and brings stability challenges during training. The policy gradient algorithms suffer from high variance when the coordination among agents is required. The details of MADRL are elaborated as follows.

#### 1) Markov Game

Markov game is a multi-agent extension of MDP. It consists of four components: a state set $S$, action sets for all the agents $A_1, A_2, \ldots, A_N$, a transition function $T(\cdot)$: $SA_1A_2\ldots A_N \rightarrow P(S)$, and reward functions for all agents $r_i$: $SA_1A_2\ldots A_N \rightarrow r$. Each agent $i$ chooses actions $A_t$ according to its local observations, and then obtains a reward that is a function of the state and action of all agents. Next, the environment reacts to all agents' action and transfers to next state. The aim of agent $i$ is to learn a policy to maximize the discounted cumulative reward $R_i = \sum_{t=0}^{T}\gamma^t r_i^t$.

#### 2) Classification of MADRL

The existing MADRL algorithms can be classified into the following groups.

1) Improved experience replay mechanism. Experience replay mechanism is a major breakthrough that enables the

combination of deep learning and RL. It helps break the correlation between training data, which is a pre-condition of the convergence of NN. However, the experience replay mechanism fails in MADRL setting since it assumes the environment to be stationary while the environment is non-stationary from any individual agent point of view. Therefore, the data sampled from the replay buffer cannot represent the current dynamics of the environment. To this end, several works try to add information to the experience tuple to help the algorithm adapt to MADRL settings [35]-[37]. Lenient-DQN-based MADRL algorithm fits into this category by assigning a leniency value to each experience tuple stored in the replay buffer. The leniency value gradually decays during training. This motivates the agent to focus on the fresh memory instead of the past experiences that no longer reflect the dynamics of the environment [37].

2) Centralized training and decentralized execution. A basic idea to guarantee a stationary environment in MADRL setting is to allow each agent know the policy of other agents. Inspired by this, [38] proposes a centralized training and decentralized execution framework based algorithm, which is named multi-agent DDPG (MADDPG). Its basic structure is shown in Fig. 6.

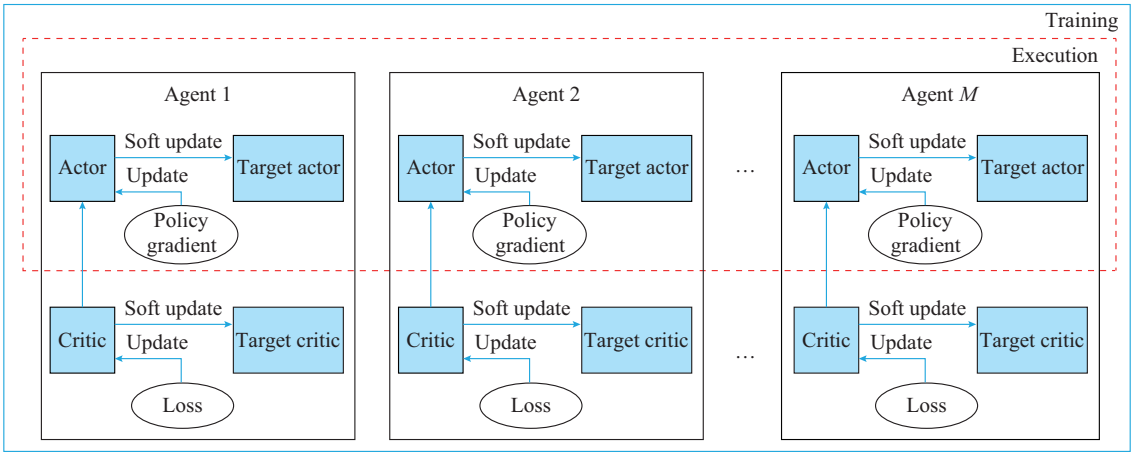


Fig. 6. Structure of MADDPG algorithms.

Each agent employs a centralized critic, which takes the global observation and actions of other agents as inputs to guarantee the Markov property. Since the global information is used by the critic during training, the actor can inform decisions based on local information when implemented in practice. Centralized training and decentralized execution framework are effective approaches to overcome the nonstationary issue in MADRL setting when off-line training can be implemented in a simulator. Attention mechanism can be further integrated with this framework to enhance the performance of the MADRL algorithm [39].

3) Recurrent network-based approaches. Recurrent NNs (RNNs) enhance the memory capability of NNs. RNNs are used in single-agent DRL to address partially observable problems and long-term credit assignment issues. Recent studies also extend RNNs to the MADRL setting to solve the challenges of partially observable Markov games [40], [41].

4) Parameter sharing. Parameter sharing is a frequent component in MADRL, which employs training a network whose parameters are shared among agents. Since the agents take different information as inputs, they can inform different decisions. This approach is proposed in [42], [43].

## III. Application in Modern Power and Energy System

Applications of RL algorithms for power and energy systems have been growing in recent years, including the optimization of smart power and energy distribution grid, flexible load demand, electricity market, and operational control and so on.

### *A. Optimization of Smart Power and Energy Distribution Grid*

#### *1) Optimization of Distribution Network*

The voltage fluctuation and power quality issues caused by the increasing penetration of DERs and electric vehicles (EVs) in distribution networks bring great challenges to the operation of the distribution network. Traditional methods such as stochastic programming and robust optimization could not effectively address highly uncertain environment. In addition, they rely heavily on the accurate parameters of the distribution system, which is difficult to obtain in practice. As a data-driven approach, DRL can provide more flexible control decisions in real time according to the latest information.

Reference [44] proposes a Monte Carlo tree search based RL method for the regulation of battery storage system to mitigate the voltage fluctuations caused by the high penetration of PV in the distribution network. Reference [45] proposes a two-timescale voltage regulation strategy combining the DQN algorithm and the physical model-based optimization. The alternating current power flow model is used for the control of smart inverters in a smaller timescale, while the DQN is to control the shunt capacitors in a larger timescale taking the long-term discounted reward value into account. These methods can inform decisions according to the latest information of distribution network in real time without the requirement of accurate physical models after training. However, the aforementioned methods deal with the con-

straints by adding penalty to the reward function. Therefore, the learned strategy may not be feasible in practice. To solve this problem, [46] proposes a volt-var control strategy of distribution network based on safe off-policy DRL algorithm. The volt-var control problem is first modeled as a constrained MDP. Then, a safe SAC algorithm is applied to solve the MDP. The proposed method explicitly models the operation constraints in the MDP, which can better satisfy the constraints, thus it is more suitable for the optimization problems with high security requirement. A Lagrangian-based DRL method is proposed in [47] for the optimization of distribution network. An approximated deterministic gradient is derived in this study instead of using the gradient provided by the critic network, which may be affected by high variance and approximation errors. Simulation results demonstrates that the proposed method can achieve similar results to the interior-point method but better capture the operation constraints than supervised learning-based approach. The accurate knowledge of physical model is still required for these methods when the reward value is caculated during training. To this end, [17] proposes a model-free voltage regulation strategy based on the surrogate model and DRL algorithm. The surrogate model is first trained in a supervised manner to capture the complex mapping from the injected power to the voltage of each node. Then, the learned surrogate model is regarded as the environment and provides the immediate reward signal to guide the optimization of the DRL algorithm. Results demonstrate that the proposed approach can achieve similar level performance to that obtained by the approaches with accurate system model. Reference [48] proposes a batch DRL based approach for the reconfiguration of distribution network. It can learn the reconfiguration strategy from the recorded data without interacting with the distribution network, thus reducing the dependence on the accurate physical models. Reference [49] proposes a batch RL-based approach for the voltage regulation by adjusting the load tap changers. A linearized power flow model is applied to estimate the voltage of each node under various tap setting conditions, which helps avoid affecting the operation of distribution network during the training process.

*2) Optimization of Microgrid*

The RL algorithms have also been applied to the optimization of microgrids in [50]-[55]. Reference [51] proposes a DRL based approach for the economic energy scheduling of microgrid embedded with renewable energies. Since the DRL algorithm does not need the system model and uncertainty information, it can be used in uncertain environment. Reference [52] proposes an RL based approach for the optimization of microgrid by utilizing the capability of battery, while in [53], a double DQN based approach is used. Reference [54] proposes a bi-level energy management strategy of microgrids. The upper level is modeled by an adaptive RL agent, the aim of which is to decide the retail price to maximize the social welfare of the entire system. Then, at the lower level, each microgrid agent solves a mixed-integer nonlinear programming problem to maximize their profits utilizing the capability of generation and storage devices. The RL agent in the upper level can automatically discover the relationship between the exchanged power at points of common coupling (PCCs) and the retail price only with information about the solar irradiance and total load demands of each microgrid. Therefore, the privacy of the customers and the microgrids are maintained. Reference [55] proposes an RL based approach for the energy management of multi-microgrids. DNN is first trained in a supervised manner to learn the behavior of multiple microgrid. Then, the Monte Carlo RL algorithm is applied to develop a near-optimal pricing strategy, with the aim to maximize the revenue and minimize the peak to average ratio at the same time. This method is suitable for the problems with great search spaces and hidden information.

*3) IES Management*

IES refers to the integrated system of energy production, supply, and marketing in the process of planning, construction, and operation. It is mainly composed of energy supply network, such as power supply, gas supply, cooling/heating network, energy exchange unit (combined cooling and heating power plant, generator set, boiler, air conditioner, heating pump, etc.), energy storage link (battery, gas storage, heat and cold storage, etc.), terminal integrated energy supply unit (microgrid) and a large number of customers, as shown in Fig. 7. Energy management of IES is challenging since it requires the coordinated regulation and control of multiple energy supply units and the collaborative optimization of supply and demand, both of which are characterized by randomness.

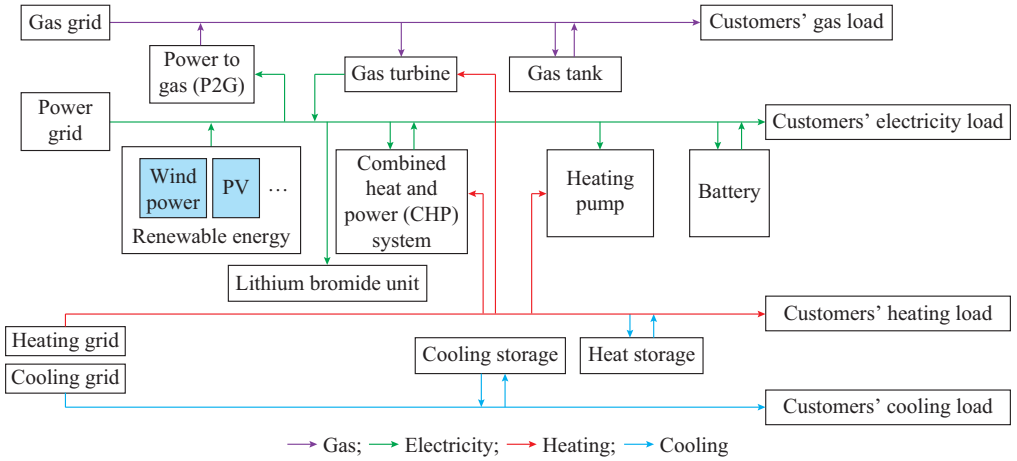


Fig. 7. Framework of IES.

Reference [56] designs an intelligent energy management system with CHP, and establishes an optimization model based on RL method aiming to minimize operation cost and carbon emission. Focusing on effectively optimizing the operation cost of IES with wind power access, [57] adopts PPO algorithm to decide wind power conversion rate. Besides, due to its remarkable self-learning advantages, PPO has better optimization performances than traditional optimization algorithms. In [58], with the aim of minimizing users' energy costs at the residential level, DDPG algorithm is applied to exploit optimal control scheme in a multi-energy system. A dynamic energy conversion and management strategy, which is based on DDPG algorithm, is proposed to smooth the net load curve while considering the economy of the system [59]. To reduce the peak load and motivate users to participate in demand response, [60] proposes an RL-based energy management strategy in electricity and natural gas network.

Employing RL-based approaches for the optimization of smart power and energy distribution grids can provide the following advantages. Firstly, they can develop near-optimal control behaviors by the continuous interaction with the environment. The learned strategy is scalable to new situations and can provide decisions in milliseconds, without resolving the problem. Therefore, they can provide more flexible control performances than pre-determined decisions when facing highly uncertain environment. Secondly, they are data-driven and reduce the dependence on accurate system model.

### *B. Demand-side Management*

The integration of renewable energy to power system must be carefully done to guarantee system security. At the same time, the users' adjustable flexible load significantly increases with the rapid development of residential smart power consumption. Demand-side management can improve the stability of power grid by changing load consumption behavior via economic incentives and increasing the flexibility of demand. As a model-free algorithm, RL can deal with the uncertainty of the environment and extract human preferences by integrating the feedback reward signal into control logic [61].

The first category to apply RL methods for demand-side management is the control of domestic hot water and heating ventilation air-conditioning devices. The objectives can be the energy cost reduction of building energy systems [62]-[64], and the increase of the energy conservation [65]-[67]. The second category is to apply RL methods to solve optimization problems of residential appliances. RL-based approaches are proposed for minimizing the cost of smart appliances [68] and shiftable loads [69]. In order to learn the optimal demand response scheduling strategy of household appliances, [70] proposes a model-free DRL method, which does not require the concrete distribution of the appliance data, electricity price, and outdoor temperature. The DNN is trained by TRPO and can effectively learn from real-time data of residential appliances. In [71], classical DRL algorithms, DQN and deep policy gradient (DPG), are used to provide a real-time optimization scheduling strategy for energy management systems. Considering the randomness of load and flexibility of PV production, [72] develops a control strategy of battery to maximize self-consumption of PV generations.

EV charging is a challenging problem owing to the randomness in the commuting behaviors of EV owner and traffic conditions, and the fluctuations of electricity price. Traditional methods rely on the forecasting information and it is difficult to obtain distribution of random variables in practice. As a model-free/data-driven approach, RL can learn the transfer probability and develop an optimal control strategy without the requirement of mathematical models.

In [73], a batch RL algorithm is proposed to reduce the charging cost of EV. Results show that the EV owner can save 10%-50% of the cost when using the proposed charging method. An adaptive energy management model based on $Q$-learning is proposed to guarantee driver's power requirement and improve the fuel economy, in which the control variable is set as the engine's torque, and the input variables are state-of-charge (SOC) of battery and the speed of generator [74]. Reference [75] explores a novel energy management strategy based on DQN for hybrid electric bus. The training results indicate that DQN-based strategy performs better than $Q$-learning and dynamic programming on time consumption and fuel economy. Similarly, compared with traditional control methods, RL-based approaches can obtain real-time decisions and achieve substantial economy savings for hybrid EVs without predefined mathematic model rules [76]-[78]. Reference [79] further improves the fuel economy of EV using DDPG algorithm, whose main idea is to consider a large amount of traffic information into the training process. Reference [80] proposes a DQN-based approach for the optimization of EV charging while [81] extends this work by integrating the DQN algorithm with the long-short term memory (LSTM) NN. To ensure that the battery operates within the allowable ranges, [82] models the charging of EV as a constrained MDP and solves it by the safe DRL.

### *C. Electricity Market*

With the increasing penetration of DERs and flexible demands, the electricity market is facing more uncertainties and complexities from both the generation and demand sides. This motivates the generation companies to design more sophisticated bidding strategy to reduce the revenue loss when participating in the liberalized electricity market. Reference [83] applies the $Q$-learning algorithm to find out the optimal bidding price in a pay-as-bid electricity market by learning from its past experience. The method reduces the dependence on the information of market clearing price and its probability distribution function. Reference [84] applies a modified RL algorithm based on the temperature variation mechanism to determine the bidding volume of the generation company. $Q$-learning and its variants are also employed for electricity market modeling in [85]-[88]. A dynamic pricing algorithm is proposed in [89] based on LSTM network and $Q$-learning algorithm. LSTM is first utilized to extract the future PV generation trends, which are then fed into the $Q$-learning algorithm for decision making. For wind

power companies, the uncertainty of wind power generation should be considered when developing the bidding strategy. Reference [90] proposes a Roth-Erev RL algorithm for the bidding strategy development of wind power plants. Note that $Q$-learning algorithms require the discretization of states and actions since they employ a look-up table to approximate the real action-value function. The discretization increases the computation burden and causes the information loss during training, which may lead to sub-optimal solutions. To this end, several improved DRL algorithms have been developed. Reference [91] proposes a DDPG-based approach for the optimization of bidding decisions of a generation company. The prioritized experience replay mechanism is also applied to enhance the capability of the algorithm. To reduce the revenue loss of wind power producer, [92] adopts the A3C algorithm for the strategic bidding of wind power producer when participating in the energy and reserve market. Reference [93] formulates the joint bidding problem of energy volume and price as an MDP, which is then solved by the DDPG algorithm. NN is used to learn a response function and extract the state transfer pattern from historical data in a supervised learning manner.

### D. Operational Control

In order to ensure the safe and stable operation of the system, different stability controllers have been developed. Typically, the parameters of controllers are tuned based on the linearized model of the system under a certain operation condition. However, the integration of more power electronics-interfaced DERs and loads makes it even more challenging.

To solve this problem, adaptive control is used for the self-tuning of the controller parameter settings. In [94], a multi-step $Q(\lambda)$ RL learning-based controller is proposed for automatic generation control (AGC) to enhance the robustness and dynamic performance of load frequency control. In [95], a $Q$-learning-based power system stabilizer (PSS) is proposed to prevent unstable low-frequency oscillations. Moreover, the data provided by the wide-area measurement system (WAMS) are used as the input signal to further enhance the control effect. Reference [96] proposes a DQN-based approach for the autonomous voltage control (AVC) of power grid. Reference [97] proposes a DQN-based power system emergency control algorithm via under-voltage load shedding and generator dynamic braking. An open-source platform for benchmarking RL algorithm for the control of power system is also developed. Reference [98] applies dueling DQN method for the autonomous topology control of power grid. Imitation learning is first applied to obtain an initial policy for the DRL agent, then dueling DQN is used to extract the optimal control strategy by interacting with the environment. After that, the domain knowledge of power systems is leveraged to increase the robustness of the learned strategy.

As mentioned above, the $Q$-learning algorithm is only suitable for the discretized action domain. To address that, in [99], the state-of-the-art DDPG, which has continuous action domain search capability, is applied as the basic method for load frequency control to minimize frequency deviation with faster response characteristics and robustness. In [100], the DDPG algorithm is also applied to train an agent to act as "grid mind" for the secure operation of power grid. After training in massive simulations, the well-trained agent masters the optimal voltage control policy, which can make autonomous voltage control strategies to support grid operators according to the real-time data from phasor measurement units (PMUs). In [101], in order to ensure the stability of the system considering the uncertainty of DERs and loads, the traditional oscillation suppression problem is formed as a faster exploration-based DRL. Then, the DDPG is introduced to solve this problem to offer flexibility and robust control to power systems. In [102], to guarantee the stability of the system with different wind speeds, a data-driven approach is proposed for the adaptive robust control of static synchronous compensator with additional damper controller (STATCOM-ADC) to address the uncertainty of the system. In [103], the A3C-based agent is proposed for the self-tuning of proportional resonance power system stabilizer (PR-PSS) to enhance the damping of the hydropower dominant system. In [104], an RL-based optimal method is proposed for the control of energy storage system (ESS) in AC-DC microgrid. Specifically, one NN is applied to act as an identifier to estimate the state of system. Subsequently, the other NN is trained to learn the optimal control policy for the ESS, which can reduce the disturbances caused by the charging and discharging of ESS.

The application summary of single-agent RL in power and energy systems is shown in Table I while those for MADRL are presented in Section III-E.

### E. Application of MADRL

Single-agent RL algorithms rely heavily on the centralized framework, which requires complete communication links and costly communication devices. With the increasing penetration of DERs and flexible loads, modern power and energy systems are becoming more complex and larger with more operation conditions and control options, which make it difficult for these methods to scale up. These issues can be effectively solved by the MADRL framework, as shown in Table II.

Two main categories are identified and reviewed here according to the implementation types.

#### 1) Independent Learner

The first category is the independent learner-based approach, which directly applies the single-agent algorithm into the multi-agent setting. Reference [105] applies multiple $Q$-learning agents to deal with the decision-making problem of multiple home appliances. Each agent models one type of appliance and all agents are trained separately. Each $Q$-learning agent aims to maximize its own reward. Reference [106] develops an MADRL-based approach for the volt-var control of three-phase unbalanced network. The actions are assigned to different agents to reduce the action dimension of each agent. All agents take the global information as input and are trained together to learn a near-optimal control strategy.

TABLE I
SUMMARY OF WORKS OF SINGLE-AGENT RL

| Reference | Field | Algorithm | Type | Objective |
|---|---|---|---|---|
| [44] | Distribution network | Monte Carlo-based RL | Others | Voltage deviation and energy consumption |
| [45] | Distribution network | DQN | Value-based | Voltage deviation |
| [46] | Distribution network | SAC | Actor-critic | Power loss and operation cost |
| [47] | Distribution network | DDPG | Actor-critic | Generation cost |
| [17] | Distribution network | DDPG | Actor-critic | Voltage deviation |
| [48] | Distribution network | Batch SAC | Actor-critic | Operation cost |
| [49] | Distribution network | Batch RL | others | Voltage deviation |
| [50] | Microgrid | A3C | Actor-critic | Operation cost |
| [51] | Microgrid | $Q$-learning | Value-based | Customers' benefit |
| [52] | Microgrid | DDPG | Actor-critic | System cost |
| [53] | Microgrid | DQN | Value-based | Operation cost |
| [54] | Microgrid | $Q$-learning | Value-based | Price signal |
| [55] | Microgrid | Monte Carlo | Value-based | Demand-side peak-to-average ratio (PAR) |
| [56] | Integrated energy system (IES) | $Q$-learning | Value-based | Operation cost and carbon emission |
| [57] | IES | PPO | Policy-based | Operation cost |
| [58] | IES | DDPG | Actor-critic | User's energy cost |
| [59] | IES | DDPG | Actor-critic | System cost and peak load shifting target |
| [60] | IES | $Q$-learning | Value-based | Peak load and customers' cost |
| [62] | Demand response | $Q$-learning | Value-based | Operation cost |
| [63] | Demand response | $Q$-learning | Value-based | Operation cost |
| [64] | Demand response | $Q$-learning | Value-based | Operation cost |
| [65] | Demand response | Fitted $Q$-iteration | Value-based | Fuel economy |
| [66] | Demand response | Fitted $Q$-iteration | Value-based | Operation cost |
| [67] | Demand response | Learning automaton | Others | Fuel economy |
| [68] | Demand response | $Q$-learning | Value-based | Customers' cost |
| [69] | Demand response | $Q$-learning | Value-based | Operating cost |
| [70] | Energy management | TRPO | Policy-based | System cost |
| [71] | Energy management | DQN | Value-based | Customers' benefit |
| [72] | Energy management | Policy iteration with $Q$ function | Value-based | Net energy cost |
| [73] | Energy management | Batch RL | Value-based | Customers' cost |
| [74] | Energy management | DQN | Value-based | Fuel economy |
| [75] | Energy management | $Q$-learning | Value-based | Fuel economy |
| [76] | Energy management | DQN | Value-based | Operation cost |
| [77] | Energy management | DQN | Value-based | Fuel economy |
| [78] | Energy management | DQN, Double DQN | Value-based | Fuel efficiency |
| [79] | Energy management | DDPG | Actor-critic | Fuel economy |
| [80] | Demand response | DQN | Value-based | Charging cost |
| [81] | Demand response | DQN | Value-based | Charging cost |
| [82] | Demand response | CPO | Value-based | Charging cost |
| [83] | Electricity market | $Q$-learning | Value-based | Supplier's profit |
| [84] | Electricity market | $Q$-learning | Value-based | Participant's benefit |
| [85] | Electricity market | $Q$-learning | Value-based | Supplier's profit |
| [88] | Electricity market | $Q$-learning | Value-based | Profit of generator |
| [89] | Electricity market | $Q$-learning | Value-based | Profit of photovaltic (PV) owner |
| [90] | Electricity market | Roth-Erev RL | Others | Profit of wind power producer |
| [91] | Electricity market | DDPG | Actor-critic | Profit of producer |
| [92] | Electricity market | A3C | Actor-critic | Profit of wind power producer |
| [93] | Electricity market | DDPG | Actor-critic | Profit of load serving entity |
| [94] | Operational control | $Q$-learning | Value-based | Frequency target |
| [95] | Operational control | $Q$-learning | Value-based | Center of angles |
| [96] | Operational control | DDPG | Actor-critic | Voltage profiles |
| [97] | Operational control | DQN | Value-based | Power system emergency control |
| [98] | Operational control | DQN | Value-based | Available transfer capabilities |
| [99] | Operational control | DDPG | Actor-critic | Frequency target |
| [100] | Operational control | DDPG | Actor-critic | Voltage profiles |
| [101] | Operational control | DDPG | Actor-critic | Speed and phase angle |
| [102] | Operational control | DDPG | Actor-critic | Damping |
| [103] | Operational control | A3C | Actor-critic | Damping |

TABLE II
SUMMARY OF WORKS OF MADRL

| Reference | Field | Algorithm | Type | Objective |
|---|---|---|---|---|
| [105] | Demand-side management | DQN | Independent learner | Electricity dissatisfaction cost |
| [106] | Distribution network | DQN | Independent learner | Voltage deviation and power loss |
| [107] | Demand-side management | $Q$-learning | Independent learner | Utility cost |
| [108] | Operational control | DQN | Independent learner | Emergency frequency control |
| [109] | Operational control | MADDPG | Centralized training and decentralized execution | Voltage deviation |
| [110] | Operational control | MADDPG | Centralized training and decentralized execution | Frequency and power deviation |
| [111] | Distribution network | MADDPG + attention | Centralized training and decentralized execution | Voltage deviation |
| [112] | Demand response | MAAC | Centralized training and decentralized execution | Energy cost |
| [113] | Distribution network | MATD3 | Centralized training and decentralized execution | Voltage deviation |

Reference [107] proposes a data-driven method for home energy management. Extreme learning algorithm is first utilized to forecast the PV generation and electricity, which are then utilized for decision making by multiple control agents modeled by $Q$-learning algorithms. Reference [108] also applies multi-$Q$-learning-based approach for the emergency frequency control.

Learning in multi-agent setting is much more complex than in single-agent cases as each agent needs to learn the dynamic of the environment as well as the policies from other agents. For each agent, the environment is nonstationary since the policies of other agents change continuously during training, leading to the violation of Markov property. Although this category of methods violates the basic assumption of RL and lacks convergence guarantees, they have actually been used in some scenarios in practice, and simulation results demonstrate that good results and better scalability can be achieved in certain circumstances.

*2) Centralized Training and Decentralized Execution*

Centralized training and decentralized execution have the general MADRL framework which employs centralized critics to guarantee the Markov property utilizing the global information during training. Reference [109] proposes an MADRL-based approach to solve the autonomous voltage control problem. The voltage control problem with several zones is first modeled as a microgrid. Then the MADDPG algorithm is applied to solve the microgrid by modeling each zone as an intelligent agent. Simulation results demonstrate its scalability for large systems. Reference [110] also applies the MADDPG algorithm for the load frequency control. The trained controller can make cooperative control decisions just based on local information. This helps reduce the dependence on highly cost communication devices. Although the centralized training and decentralized execution mechanism help solve the nonstationary problem in multi-agent setting, it is still challenging to address problems with large populations. To this end, [111] proposes an attention-based MADDPG for the optimization of distribution network. The attention mechanism helps each agent attend to the specific information that is mostly related to its immediate reward. Thus, it is suitable for problems with large populations. Reference [112] also applies the attention-based MADRL algorithm for the optimization of energy system with heating ventilation air-conditioning devices in multiple commercial buildings. However, the function approximation errors persists in DDGP algorithm and may lead to suboptimal policies. To this end, [113] develops the multi-agent twin delayed DDPG (TD3) algorithm, which also adopts the centralized training and decentralized execution framework while modeling each control subject as a TD3 agent. Simulation results demonstrate that the agents can make real-time near-optimal decisions just based on local information.

## IV. CONCLUSION AND FUTURE WORK

The increasing complexity and uncertainty in modern power and energy systems, as well as the wide-area deployment of advanced sensors make the ML-based approach a promising alternative for power system operation and control. This paper conducts a comprehensive review of RL algorithms and their applications in power and energy systems. A review of widely accepted algorithms in RL, DRL, and MADRL is first provided. Then, the applications of RL algorithms in power and energy systems are investigated in detail, including the optimization of distribution networks and microgrid, energy management, electricity market, demand response, and operation control. Several applications of MADRL are presented as well.

Although numerous applications of RL in modern power and energy systems have been studied, there are still many interesting problems worth further studies, including but are not limited to the follows.

1) Since the power and energy systems have a high requirement for safety, the physical constraints should be better handled when building the RL model instead of directly adding soft constraints to the reward function. Safe RL is a suitable way to deal with the optimization and control problems by solving a constrained MDP. Other ways that can embed the physical knowledge in the RL model may also improve the reliability and motivate the real-world implementation.

2) The offline training relies on the accurate physical model while the online training may affect the operation of power and energy systems. Batch RL and surrogate model are two ways to reduce the dependence on physical model without impacting the operation of system. However, they require a certain amount of historical data. Transfer learning may be another promising alternative by training the RL

model offline and transferring the learned control strategy to real-world environments with few-shot recorded samples and iterations.

3) Modern power and energy systems are becoming more complex and larger with more operation conditions and control options. Single-agent RL algorithm adopts centralized framework that relies heavily on the complete communication links, and thus is incapable of dealing with communication delay and scaling up to large systems. MADRL can partially mitigate this issue by adopting a centralized training, decentralized execution framework. However, existing MADRL algorithms face great challenges when dealing with very large systems that require large populations. Further research may apply advanced MADRL algorithms with novel population scaling mechanisms to enable the RL method to scale up to very large systems.

4) A lot of control and optimization problems in power and energy systems have typical hierarchical structures, as well as the decision-making process of human being. Hierarchical framework can reduce the deployment cost of complete communication devices of centralized control and avoid the isolation issue of local control, and thus it is another promising way for the control of large systems. Owing to the complexity of hierarch structure and the lack of a general hierarchical framework, applications of RL for hierarchical control are rare in power and energy systems. Future research may apply RL-based hierarchical control framework for large systems.

5) With the increased integration power electronic devices, DERs and flexible loads, the complexities and uncertainties are growing in modern power and energy systems. Classical offline training and online execution manner of RL is incapable of dealing with the continuously generated unmolded dynamics. Meta-learning and continuous learning can be integrated with the RL algorithm to achieve the life-long learning ability. This helps the continuous transformation of online data into powerful knowledge, which can successively enhance the control behavior of the RL agent. Therefore, the robustness and adaptability to unmolded system dynamics can be enhanced, and the trainning time can be shorten in complex scenarios.

## REFERENCES


[1] R. Detchon and R. Van Leeuwen, "Policy: bring sustainable energy to the developing world," *Nature*, vol. 508, no. 7496, pp. 309-311, Apr. 2014.

[2] B. Kroposki, "Integrating high levels of variable renewable energy into electric power systems," *Journal of Modern Power Systems and Clean Energy*, vol. 5, no. 6, pp. 831-837, Nov. 2017.

[3] J. Zhu, E. Zhuang, J. Fu *et al*., "A framework-based approach to utility big data analytics," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1-8, Aug. 2015.

[4] D. Cao, J. Li, D. Cai *et al*., "Design and application of big data platform architecture for typical scenarios of power system," in *Proceedings of 2018 IEEE PES General Meeting* (PESGM), Portland, USA, Aug. 2018, pp. 1-5.

[5] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 237-285, Apr. 1996.

[6] Y. Xu, Z. Dong, R. Zhang *et al*., "Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4398-4408, Feb. 2017.

[7] P. Li, C. Zhang, Z. Wu *et al*., "Distributed adaptive robust voltage/var control with network partition in active distribution networks," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2245-2256, Oct. 2019.

[8] B. Zhao, Z. Xu, C. Xu *et al*., "Network partition-based zonal voltage control for distribution networks with distributed PV systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4087-4098, Jan. 2017.

[9] F. L. Pagola, I. J. Perez-Arriaga, and G. C. Verghese, "On sensitivities, residues and participations: applications to oscillatory stability analysis and control," *IEEE Transactions on Power Systems*, vol. 4, no. 1, pp. 278-285, Mar. 1989.

[10] C. Chung, L. Wang, F. Howell *et al.,* "Generation rescheduling methods to improve power transfer capability constrained by small-signal stability," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 524-530, Mar. 2004.

[11] Z. Bouchama and M. Harmas, "Optimal robust adaptive fuzzy synergetic power system stabilizer design," *Electric Power Systems Research*, vol. 83, no. 1, pp. 170-175, Feb. 2012.

[12] S. Das and I. Pan, "On the mixed $H_\infty/H_2$ loop-shaping tradeoffs in fractional-order control of the AVR system," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 1982-1991, Nov. 2013.

[13] D. Ke, F. Shen, C. Chung *et al*., "Application of information gap decision theory to the design of robust wide-area power system stabilizers considering uncertainties of wind power," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 805-817, Apr. 2018.

[14] P. Zhao, W. Yao, S. Wang *et al*., "Decentralized nonlinear synergetic power system stabilizers design for power system stability enhancement," *International Transaction on Electrical Energy Systems*, vol. 24, no. 9, pp. 1356-1368, Sept. 2014.

[15] M. J. Morshed and A. Fekih, "A probabilistic robust coordinated approach to stabilize power oscillations in DFIG-based power systems," *IEEE Transactions on Industrial Info*rmatics, vol. 15, no. 10, pp. 5599-5612, Oct. 2019.

[16] Z. Ni, Y. Tang, X. Sui *et al*., "An adaptive neuro-control approach for multi-machine power systems," *International Journal of Electrical Power & Energy Systems*, vol. 75, pp. 108-116, Feb. 2016.

[17] D. Cao, J. Zhao, W. Hu *et al*. (2020, Jun.). Model-free voltage regulation of unbalanced distribution network based on surrogate model and deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/2006.13992.

[18] Y. Gao, W. Wang, J. Shi *et al*., "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357-5369, Nov. 2020.

[19] Y. Gao, R. Zhou, H. Wang *et al*., "Study on an average reward reinforcement learning algorithm," *Chinese Journal of Computers*, vol. 30, no. 8, pp. 1372-1378, Aug. 2007.

[20] E. Ipek, O. Mutlu, J. F. Mart *et al*., "Self-optimizing memory controllers: A reinforcement learning approach," *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, pp. 39-50, Jul. 2008.

[21] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: a survey," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238-1274, Jan. 2013.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*. Cambridge: MIT Press, 1998.

[24] P. Hernandezleal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-agent Systems*, vol. 33, no. 6, pp. 750-797, Oct. 2019.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[26] V. Mnih, K. Kavukcuoglu, D. Silver *et al*. (2013, Dec.). Playing Atari with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1312.5602

[27] H. Van Hasselt, A. Guez, and D. Silver. (2015, Sept.). Deep reinforcement learning with double *Q*-learning. [Online]. Available: https://arxiv.org/abs/1509.06461v1

[28] Z. Wang, T. Schaul, M. Hessel *et al*., "Dueling network architectures for deep reinforcement learning," in *Proceedings of International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 1995-2003.

[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al*., "Continuous control with deep reinforcement learning," in *Proceedings of International Conference on Learning Representation (ICLR)*, San Diego, USA, May 2015, pp. 1-14.

[30] D. Silver, G. Lever, N. Heess *et al*., "Deterministic policy gradient algorithms," in *Proceedings of International Conference on Machine Learning*, Beijing, China, Jun. 2014, pp. 387-395.

[31] V. Mnih, A. P. Badia, M. Mirza *et al*. (2016, Jun.). Asynchronous

methods for deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1602.01783
[32] T. Haarnoja, A. Zhou, P. Abbeel *et al*. (2018, Jan.). Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. [Online]. Available: https://arxiv.org/abs/1801.01290
[33] J. Schulman, S. Levine, P. Abbeel *et al*., "Trust region policy optimization," in *Proceedings of International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 1889-1897.
[34] J. Schulman, F. Wolski, P. Dhariwal *et al*. (2017, Jul.). Proximal policy optimization algorithms. [Online]. Available: https://arxiv. org/abs/1707.06347
[35] S. Omidsha_Ei, J. Pazis, C. Amato *et al*., "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, Aug. 2017, pp. 2681-2690.
[36] J. N. Foerster, N. Nardelli, G. Farquhar *et al*., "Stabilizing experience replay for deep multi-agent reinforcement learning," in *Proceedings of International Conference on Machine Learning*, Sydney, Australia, Aug. 2017, pp. 1-10.
[37] G. Palmer, K. Tuyls, D. Bloembergen *et al*. (2017, Jul.). Lenient multi-agent deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1707.04402
[38] R. Lowe, Y. Wu, A. Tamar *et al*., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, USA, Jun. 2017, pp. 6379-6390.
[39] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *Proceedings of International Conference on Machine Learning*, Stockholm, Sweden, Oct. 2018, pp. 2961-2970.
[40] Z. Hong, S. Su, T. Shan *et al*., "A deep policy inference *Q*-network for multi-agent systems," in *Proceedings of International Conference on Autonomous Agents and Multiagent Systems*, Sao Paulo, Brazil, Nov. 2017, pp. 1388-1396.
[41] M. Jaderberg, W. M. Czarnecki, D. Iain *et al*., "Human-level performance in 3D multiplayer games with population-based reinforcement learning," *Science*, vol. 364, no. 6443, pp. 859-865, May 2019.
[42] J. N. Foerster, Y. M. Assael, N. De Freitas *et al*., "Learning to communicate with deep multi-agent reinforcement learning," in *Proceedings of Advances in Neural Information Processing Systems*, Barcelona, Spain, May 2016, pp. 2145-2153.
[43] J. K. Gupta, M. Egorov, and M. J. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *Proceedings of International Conference on Autonomous Agents and Multiagent Systems*, Sao Paulo, Brazil, Nov. 2017, pp. 66-83.
[44] M. Al-Saffar and P. Musilek, "Reinforcement learning based distributed BESS management for mitigating overvoltage issues in systems with high PV penetration," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 2980-2994, Feb. 2020.
[45] Q. Yang, G. Wang, A. Sadeghi *et al*., "Two-timescale voltage control in distribution grids using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2313-2323, Nov. 2019.
[46] W. Wang, N. Yu, Y. Gao *et al*., "Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3008-3018, Dec. 2019.
[47] Z. Yan and Y. Xu, "Real-time optimal power flow: a Lagrangian based deep reinforcement learning approach," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3270-3273, Apr. 2020.
[48] Y. Gao, W. Wang, J. Shi *et al*., "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357-5369, Nov. 2020.
[49] H. Xu, A. D. Domínguez-García, and P. W. Sauer, "Optimal tap setting of voltage regulation transformers using batch reinforcement learning," *IEEE Transactions on Power Systems*, vol. 35, no. 3, pp. 1990-2001, Oct. 2019.
[50] V. François-Lavet, D. Taralla, and D. Ernst, "Deep reinforcement learning solutions for energy microgrids management," in *Proceedings of European Workshop on Reinforcement Learning (EWRL)*, Barcelona, Spain, Nov. 2016, pp. 1-7.
[51] E. Kuznetsova, Y. Li, C. Ruiz, *et al*., "Reinforcement learning for microgrid energy management," *Energy*, vol. 59, pp. 133-146, May 2013.
[52] X. Yang, Y. Wang, H. He *et al*., "Deep reinforcement learning for economic energy scheduling in data center microgrids," in *Proceedings of 2019 IEEE PES General Meeting (PESGM)*, Atlanta, USA, Aug. 2019, pp. 1-5.
[53] V. H. Bui, A. Hussain, and H. M. Kim, "Double deep *Q*-learning-based distributed operation of battery energy storage system considering uncertainties," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 457-469, Jun. 2019.
[54] Q. Zhang, K. Dehghanpour, Z. Wang *et al*., "A learning-based power management method for networked microgrids under incomplete information," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1193-1204, Aug. 2020.
[55] Y. Du and F. Li, "Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1066-1076, Jul. 2019.
[56] Q. Sun, D. Wang, D. Ma *et al*., "Multi-objective energy management for we-energy in Energy Internet using reinforcement learning," in *Proceedings of 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, USA, Nov. 2017, pp. 1-6.
[57] B. Zhang, W. Hu, D. Cao *et al*., "Deep reinforcement learning-based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy," *Energy Conversion and Management*, vol. 202, p. 112199, Dec. 2019.
[58] Y. Ye, D. Qiu, X. Wu *et al*., "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068-3082, Feb. 2020.
[59] B. Zhang, W. Hu, J. Li *et al*., "Dynamic energy conversion and management strategy for an integrated electricity and natural gas system with renewable energy: deep reinforcement learning approach," *Energy Conversion and Management*, vol. 220, p. 113063, Sept. 2020.
[60] A. Sheikhi, M. Rayati, and A. M. Ranjbar, "Demand side management for a residential customer in multi-energy systems," *Sustainable Cities and Society*, vol. 22, pp. 63-77, Jan. 2016.
[61] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: a review of algorithms and modeling techniques," *Applied Energy*, vol. 235, pp. 1072-1089, Feb. 2019.
[62] G. P. Henze and S. Liu, "Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory – part 1: theoretical foundation." *Energy and Buildings*, vol. 38, no. 2, pp. 142-147, Feb. 2006.
[63] G. P. Henze and S. Liu, "Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory – part 2: results and analysis," *Energy and Buildings*, vol. 38, no. 2, pp. 148-161, Feb. 2006.
[64] G. P. Henze and S. Liu, "Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory," *Journal of Solar Energy Engineering*, vol. 129, no. 2, pp. 215-225, May 2007.
[65] J. Vázquez-Canteli, J. Kampf, and Z. Nagy, "Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted *Q*-iteration," *Energy Procedia*, vol. 122, pp. 415-420, Sept. 2017.
[66] F. Ruelens, B. J. Claessens, S. Vandael *et al*., "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149-2159, Feb. 2016.
[67] H. Kazmi, F. Mehmood, S. Lodeweyckx *et al*., "Gigawatt-hour scale savings on a budget of zero: deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159-168, Dec. 2017.
[68] Y. Liang, L. He, X. Cao *et al*., "Stochastic control for smart grid users with flexible demand," *IEEE Transaction on Smart Grid*, vol. 4, no. 4, pp. 2296-2308, Dec. 2013.
[69] Y. Liu, C. Yuen, N. U. Hassan *et al*., "Electricity cost minimization for a microgrid with distributed energy resource under different information availability," *IEEE Transaction on Industrial Electronics*, vol. 62, no. 4, pp. 2571-2583, Jan. 2014.
[70] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Transaction on Smart Grid*, vol. 11, no. 5, pp. 4144-4154, Mar. 2020.
[71] E. Mocanu, D. C. Mocanu, P. H. Nguyen *et al*., "On-line building energy optimization using deep reinforcement learning," *IEEE Transaction on Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2017.
[72] B. V. Mbuwir, F. Spiessens, and G. Deconinck, "Self-learning agent for battery energy management in a residential microgrid," in *Proceedings of 2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Sarajevo, Bosnia, Oct. 2018, pp. 1-6.
[73] A. Chis, J. Lunden, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price," *IEEE Transaction on Vehicular Technology*, vol. 66, no. 5, pp. 3674-3684, Jan. 2016.
[74] J. Wu, H. He, J. Peng *et al*., "Continuous reinforcement learning of en-

ergy management with deep *Q* network for a power split hybrid electric bus," *Applied Energy*, vol. 222, pp. 799-811, Jul. 2018.
[75] T. Liu, Y. Zou, D. Liu *et al*., "Reinforcement learning of adaptive energy management with transition probability for a hybrid electric tracked vehicle," *IEEE Transaction on Industrial Electronics*, vol. 62, no. 12, pp. 7837-7846, Dec. 2015.
[76] Y. Hu, W. Li, K. Xu *et al*., "Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning," *Applied Sciences*, vol. 8, no. 2, pp. 187-202, Jan. 2018.
[77] X. Qi, Y. Luo, G. Wu *et al*., "Deep reinforcement learning-based vehicle energy efficiency autonomous learning system," in *Proceedings of 2017 IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, USA, Jun. 2017, pp. 1228-1233.
[78] X. Qi, Y. Luo, G. Wu *et al*., "Deep reinforcement learning enabled self-learning control for energy efficient driving," *Transportation Research Part C: Emerging Technologies*, vol. 99, pp. 67-81, Feb. 2019.
[79] Y. Wu, H. Tan, J. Peng *et al*., "Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus," *Applied Energy*, vol. 247, pp. 454-466, Aug. 2019.
[80] Z. Wan, H. Li, H. He *et al*., "A data-driven approach for real-time residential EV charging management," in *Proceedings of 2018 IEEE Power & Energy Society General Meeting (PESGM)*, Portland, USA, Aug. 2018, pp. 1-5.
[81] Z. Wan, H. Li, H. He *et al*., "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Nov. 2018.
[82] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427-2439, Nov. 2019.
[83] M. Rahimiyan and H. R. Mashhadi, "Supplier's optimal bidding strategy in electricity pay-as-bid auction: comparison of the *Q*-learning and a model-based approach," *Electric Power Systems Research*, vol. 78, no. 1, pp. 165-175, Jan. 2008.
[84] M. B. Naghibi-Sistani, M. Akbarzadeh-Tootoonchi, M. J. D. Bayaz *et al*., "Application of *Q*-learning with temperature variation for bidding strategies in market based power systems," *Energy Conversion and Management*, vol. 47, no. 11, pp. 1529-1538, Jan. 2006.
[85] G. Xiong, T. Hashiyama, and S. Okuma, "An electricity supplier bidding strategy through *Q*-learning," in *Proceedings of IEEE Power Engineering Society Summer Meeting*, Chicago, USA, vol. 3, Aug. 2002, pp. 1516-1521.
[86] H. Song, C. Liu, J. Lawarree *et al*., "Optimal electricity supply bidding by Markov decision process," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 618-624, Jun. 2000.
[87] V. Nanduri and T. K. Das, "A reinforcement learning model to assess market power under auction-based energy pricing," *IEEE Transactions on Power Systems*, vol. 22, no. 1, pp. 85-95, Mar. 2007.
[88] A. C. Tellidou and A. G. Bakirtzis, "Agent-based analysis of capacity with holding and tacit collusion in electricity markets," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1735-1742, Dec. 2007.
[89] X. Xu, Y. Xu, J. Li *et al.*, "Data-driven game-based pricing for sharing rooftop photovoltaic generation and energy storage in the residential building cluster under uncertainties," *IEEE Transactions on Industrial Informatics.* doi: 10.1109/TII.2020.3016336
[90] G. Li and J. Shi, "Agent-based modeling for trading wind power with uncertainty in the day-ahead wholesale electricity markets of single-sided auctions," *Applied Energy*, vol. 99, pp. 13-22, Nov. 2012.
[91] Y. Ye, D. Qiu, M. Sun *et al*., "Deep reinforcement learning for strategic bidding in electricity markets," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1343-1355, Aug. 2019.
[92] D. Cao, W. Hu, X. Xu *et al*., "Bidding strategy for trading wind energy and purchasing reserve of wind power producer – a DRL based approach," *International Journal of Electrical Power & Energy Systems*, vol. 117, p. 105648, May 2020.
[93] H. Xu, H. Sun, D. Nikovski *et al*., "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6366-6375, Mar. 2019.
[94] T. Yu, B. Zhou, K. W. Chan *et al*., "Stochastic optimal relaxed automatic generation control in non-Markov environment based on multi-step $Q(\lambda)$ learning," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1272 -1282, Aug. 2011.
[95] R. Hadidi and B. Jeyasurya, "Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 489-497, Mar. 2013.
[96] R. Diao, Z. Wang, D. Shi *et al*., "Autonomous voltage control for grid operation using deep reinforcement learning," in *Proceedings of IEEE PES General Meeting*, Atlanta, USA, Aug. 2019, pp. 1-5.
[97] Q. Huang, R. Huang, W. Hao *et al*., "Adaptive power system emergency control using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1171-1182, Aug. 2019.
[98] T. Lan, J. Duan, B. Zhang *et al*. (2019, Nov.). AI-based autonomous line flow control via topology adjustment for maximizing time-series ATCs. [Online]. Available: https://arxiv.org/abs/1911.04263
[99] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: a deep reinforcement learning method with continuous action search," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1653-1656, Nov. 2018.
[100] J. Duan, D. Shi, R. Diao *et al*., "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 814-817, Sept. 2019.
[101] Y. Hashmy, Z. Yu, D. Shi *et al*., "Wide-area measurement system-based low frequency oscillation damping control through reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5072-5083, Nov. 2020.
[102] G. Zhang, W. Hu, D. Cao *et al*., "A data-driven approach for designing statcom additional damping controller for wind farms," *International Journal of Electrical Power & Energy Systems*, vol. 117, p. 105620, May 2020.
[103] G. Zhang, W. Hu, D. Cao *et al*., "Deep reinforcement learning-based approach for proportional resonance power system stabilizer to prevent ultra-low-frequency oscillations," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5260-5272, Nov. 2020.
[104] J. Duan, Z. Yi, D. Shi *et al*., "Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC-DC microgrids," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5355-5364, Jan. 2019.
[105] R. Lu, S. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transaction on Smart Grid*, vol. 10, no. 6, pp. 6629-6639, Apr. 2019.
[106] Y. Zhang, X. Wang, J. Wang *et al*., "Deep reinforcement learning based volt-var optimization in smart distribution systems," *IEEE Transactions on Smart Grid*. doi: 10.1109/TSG.2020.3010130
[107] X. Xu, Y. Jia, Y. Xu *et al*., "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Transaction on Smart Grid*, vol. 11, no. 4, pp. 3201-3211, Jul. 2020.
[108] C. Chen, M. Cui, F. Li *et al*., "Model-free emergency frequency control based on reinforcement learning," *IEEE Transactions on Industrial Informatics*. doi: 10.1109/TII.2020.3001095
[109] S. Wang, J. Duan, D. Shi *et al*., "A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning," *IEEE Transactions on Power Systems,* vol. 35, no. 6, pp. 4644-4654, Nov. 2020.
[110] Z. Yan and Y. Xu, "A multi-agent deep reinforcement learning method for cooperative load frequency control of multi-area power systems," *IEEE Transactions on Power System*s, vol. 35, no. 6, pp. 4599-4608, Nov. 2020.
[111] D. Cao, W. Hu, J. Zhao *et al*., "A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4120-4123, Sept. 2020.
[112] L. Yu, Y. Sun, Z. Xu *et al*., "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Transactions on Smart Grid.* doi: 10.1109/TSG.2020.3011739
[113] D. Cao, J. Zhao, W. Hu *et al*. (2020, May). Distributed voltage regulation of active distribution system based on enhanced multi-agent deep reinforcement learning. [Online]. Available: https://arxiv. org/abs/2006.00546

**Di Cao** is currently pursuing the Ph.D. degree in control science and engineering with University of Electronic Science and Technology of China, Chengdu, China. His research interests include optimization of distribution network and application of machine learning algorithms in power system.

**Weihao Hu** received the B.Eng. and M.Sc. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 2012. He is currently a Full Professor and the Director of the Institute of Smart Power and Energy Systems, University of Electronics Science and Technology of China, Chengdu, China. His research interests include artifi-

cial intelligence in modern power system and renewable power generation.

**Junbo Zhao** received the Ph. D. degree from the Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Falls Church, USA, in 2018. Now he is an Assistant Professor with Mississippi State University, Starkville, USA. His research interests include power system modeling, real-time monitoring, dynamics and cyber security, big data analytic, and robust statistical signal processing.

**Guozhou Zhang** received the B. S. degree from Chongqing University of Technology, Chongqing, China, in 2016, the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2019, where he is currently pursuing the Ph. D. degree in control science and engineering. His research interest includes power system analysis and control.

**Bin Zhang** received the B.S. degree from Hohai University, Nanjing, China, in 2017. He is currently pursuing the M.S. degree in University of Electronic Science and Technology of China, Chengdu, China. His research interest is optimization of hybrid energy system.

**Zhou Liu** received the B.Eng. and M.Sc. degrees in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004 and 2007, respectively, and the Ph.D. degree in energy technology from the Department of Energy Technology, Aalborg University, Aalborg, Denmark, in 2013. He is currently with the Department of Energy Technology, Aalborg University as an Assistant Professor. His research interests include power system analysis and digital simulation, wide-area protection and control, wind power integration and power substation automation, high-votage direct current (HVDC) circuit breaker and protection.

**Zhe Chen** received the B.Eng. and M.Sc. degrees from the Northeast China Institute of Electric Power Engineering, Jilin, China, and the Ph. D. degree from the University of Durham, Durham, UK. He is a Full Professor with the Department of Energy Technology, Aalborg University, Aalborg, Denmark. His research areas include power systems, power electronics and electric machines, and his main current research interests include wind energy and modern power systems.

**Frede Blaabjerg** received the Ph. D. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1995. He was with ABB-Scandia, Randers, Denmark, from 1987 to 1988. He became an Assistant Professor, in 1992, an Associate Professor in 1996, and a Full Professor of power electronics and drives in 1998. In 2017, he became a Villum Investigator. He is Honoris Causa at University Politehnica Timisoara, Timisoara, Romania, and Tallinn Technical University, Tallinn, Estonia. His current research interests include power electronics and its applications such as in wind turbines, PV systems, reliability, harmonics and adjustable speed drives.